A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients

Norul Hidayah Ibrahim[1], Aida Mustapha[2], Rozilah Rosli[3], Nurdhiya Hazwani Helmee[4]

Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia (UPM),
43400 UPM Serdang, Selangor, Malaysia
[1] norulhidayah@gmail.com
[2] aida_m@upm.edu.my
[3] rozirose77@gmail.com
[4] deeya_wanie@hotmail.com

***Abstract*— Hybrid models in data mining have recently gained attention including in the study of medical research. Various studies in this domain using hybrid models have shown different results. This paper presents the new hybrid model by exploring Agglomerative Hierarchical Clustering and Decision Tree Classifier on Pima Indians Diabetes dataset. The experiments compared performance accuracy of the Decision Tree Classifier against the same classifier augmented with Hierarchical Clustering. Results showed that the hybrid model achieved higher accuracy with 80.8% as compared to 76.9% of the standard model. This is a promising result for adoption of hierarchical clustering in a rule-based classifier.**

Keyword- Medical Data Mining, Hierarchical Clustering, Classification, Decision Tree, Hybrid Model

I. INTRODUCTION

Diabetes is a chronic disease that begin with failure of pancreas to produce sufficient insulin or when the body is not able to effectively use the insulin it produces. The internal changes prompt to an increased concentration of glucose in the blood, namely hyperglycemia. Hyperglycemia is a condition where there is high blood glucose in a diabetic patient. Again, it happens when insulin inside the body is not enough to manage the blood sugar level. Hyperglycemia can be caused by either Type I or Type II diabetes. According to the World Health Organization, Type I diabetes is known as insulin-dependent or childhood-onset diabetes. It is characterized by lack of insulin production. Meanwhile, Type II diabetes is known as non-insulin-dependent or adult-onset diabetes, which is caused by ineffective use of insulin by human body. This will result in excess body weight and physical inactivity.

Obesity, high cholesterol, and high blood pressure are the major risk factors that trigger diabetes. The risks are higher as when the age increases. Also, pregnant women with a condition of gestational diabetes are more likely to get permanent diabetes in their life. In addition, long term complications of this disease include heart attacks, strokes, blindness, and many others [1], [2]. World Health Organization (WHO) manifested a rising number of diabetic cases. This trend is anticipated to grow in the next couple of decades. In the National Diabetic Fact Sheet 2011, it is predicted 2% to 10% of pregnant women will develop gestational diabetes and from this group, they are more likely to get Type II diabetes later in their life. By 2030, the number of cases worldwide is expected to rise to 8.4 billion from 7.0 billion in 2010 [3].

Pima Indians Diabetes dataset [4] has been widely studied in the area of Data Mining, which is an interdisciplinary field merging from statistics, machine learning, information science, visualization and other disciplines [5]. Data mining capitalizes on big datasets to obtain unknown patterns in the data by various tasks such as association rule mining, classification or prediction, and clustering. Association analysis is mainly used to mine various kinds of association or correlation rules by finding frequent patterns, associations, correlations or causal structures among set of items in databases or any information repositories. Classification predicts class labels, while prediction estimates future value of an attribute. Clustering groups the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [6]. In contrast to clustering, classification is a type of supervised learning because the data classes are known beforehand. Recent work in classification focus on hybrid models by incorporating clustering prior to the classification task.

This paper proposes a hybrid classification model to predict diabetic cases using Agglomerative Hierarchical algorithm for clustering with a Decision Tree-based classifier validated on the standard Pima Indians Diabetic dataset. Implementation of the proposed model will be with the Waikato Environment for Knowledge Analysis

(WEKA) [7]. The remainder of this paper is organized as follows. Section 2 highlights related work on hybrid classification models. Section 3 presents the proposed hybrid classification model and describes the methods for the experiment. Section 4 discusses the experimental results, and finally Section 5 concludes with some direction of future work.

II. Related Work

Hybrid means a thing made by combining two different elements. In this paper, hybrid classification model refers to a combination of two data mining tasks, which are clustering and classification in effort to obtain higher accuracy result. Previously, hybrid classification models have been applied to predict patients who have higher risk in having diabetes by looking at the patients' profiles [1], [2], [8]. This information is useful to help overcome a predictable diabetic patient. Hybrid Prediction Model (HPM) by [8] uses Simple *K*-means clustering algorithm together with a C4.5 classifier. This model is trained on patients' characteristics and measurements in order to inquire on how diabetes occurrences are affected. By using *k*-fold cross validation method in Decision Tree C4.5 Classifier, the accuracy increased to 92.4% compared to C4.5 algorithm alone, which only fared between the range of 59.4% to 84.1% in classification accuracy.

Subsequent work by [6] introduces a hybrid model using *K*-Nearest Neighbour, Expectation Maximization and Genetic Algorithms. To ensure the successful result, this work removes data that are difficult to learn [6]. In 2012, patients who are in high risk of having diabetic have also been compartmentalized by a hybrid classification model, using *K*-means clustering with *K*-Nearest Neighbour classifier [2]. In the research, the hybrid model deals with 392 records with no missing values, whereby 130 records are tested positive cases and 262 records are tested negative. Input for these tasks are number of times pregnant, plasma glucose concentration of 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function and age. The results showed that the hybrid model has raised the accuracy of *K*-Nearest Neighbour up to 96.7% for diabetic dataset.

Similar but latest hybrid model consists of *K*-Means clustering and Decision Tree C4.5 classifier. This model, again, purposely proposed to compartmentalize patients who are in high risk of having diabetic and patients who are not by using different algorithms [1]. Using the same dataset, which consists of 392 cases with no missing values, the proposed cascaded model obtained the classification accuracy of 93.3 % when compared to accuracy of 73.6 % using C4.5 classifier alone.

A. Clustering

In this paper, we propose hierarchical clustering as a method of cluster analysis which follows to build a hierarchy of clusters. Hierarchical cluster analysis (or hierarchical clustering) use distance matrix for clustering. It is a general approach to cluster analysis, in which the objective is to group together objects or records that are most similar to one another. This method does not require initial specification of the number of clusters *k* as an input. The calculation of distance measures between objects and cluster is repeated once the objects begin to be grouped into clusters. The outcome is represented graphically as a dendrogram where it shows how the clusters are merged hierarchically. The most common algorithm for hierarchical clustering is agglomerative hierarchical clustering where each data point forms a cluster in the beginning. The merging begins when the merged nodes or clusters have least dissimilarity. Eventually all nodes belong to the same cluster.

Agglomerative hierarchical clustering algorithm consists of two stages where it joins the closest clusters that are most similar to one other. During the initial stage, each cluster has only one object and the cluster will grow to include one or more of the closest objects. However, the method has different ways of defining the similarity between clusters. Agglomerative approach treats each feature as one independent cluster and then successively merges pairs of clusters until all clusters have been merged into a single cluster. Agglomerative hierarchical clustering requires no prior information regarding the number of clusters required, are easy to implement, and provide best results [8]. Agglomerative hierarchical clustering consist of four types such as single linkage clustering, complete linkage clustering, average linkage clustering, and Cobweb (Available at http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf).

B. Classification

It has been mentioned earlier that classification is a supervised learning algorithm because the class labels are known for the examples used to build the classifier. In classification, the general task is to assign class labels to a set of unclassified new data. Decision Tree is one of the supervised learning algorithms that use the approach of tree generation or recursive top-down structure [9]. There are many specific Decision Tree algorithms and the most widely used to date is the Iterative Dichotomiser 3 (ID3) [9]. The improvement of ID3 is the C4.5 algorithm [10], which is known as J48 algorithm in WEKA data mining tool [11].

A number of studies have reported the performance of Decision Tree-based classifier validated on Pima Indians Diabetes dataset. Some presented result on mining with Decision Tree C4.5 alone and some study applied Decision Tree with other data mining techniques. Table 1 shows accuracies of the proposed algorithms

and other classification methods on Pima Indians Diabetes dataset [1], [12]. There is also a growing interest on comparative study among supervised learning algorithms including Decision Tree that deals with Pima Indians Dataset [13].

TABLE I
Comparison of Classification Accuracy between Proposed and Existing Methods on PIDD

Method	**Accuracy (%)**
Hybrid Model (Hierarchical Clustering + Decision Tree J48)	80.77
Proposed Model k-means+DT continuous data)	93.33
K-means +KNN, k=5	96.68
ANN+FNN	82.4
Log disc	77.7
IncNet	77.6
DIPOL92	77.6
Linear Discr. Anal.	77.5–77.2
SMART	76.8
GTO DT (5 · CV)	76.8
kNN, k = 23, Manh, raw	76.7 ± 4.0
kNN, k = 1:25, Manh, raw	76.6 ± 3.4
ASI	76.6
Fisher discr. analysis	76.5
MLP + BP	76.4
MLP + BP	75.8 ± 6.2
LVQ	75.8
LFC	75.8
RBF	75.7
NB	75.5–73.8
kNN, k = 22, Manh	75.5
MML	75.5 ± 6.3
SNB	75.4
BP	75.2
SSV DT	75.0 ± 3.6
kNN, k = 18, Euclid, raw	74.8 ± 4.8
CART DT	74.7 ± 5.4
CART DT	74.5
DB-CART	74.4
ASR	74.3
SSV DT	73.7 ± 4.7
C4.5 DT	73.0
C4.5 DT	72.7 ± 6.6
Bayes	72.2 ± 6.9
C4.5 (5 · CV)	72.0
CART	72.8
Kohonen	72.7
kNN	71.9
ID3	71.7 ± 6.6
IB3	71.7 ± 5.0
IB1	70.4 ± 6.2
kNN, k = 1, Euclides, raw	69.4 ± 4.4
kNN	67.6
C4.5 rules	67.0 ± 2.9
OCN2	65.1 ± 1.1
QDA	59.5

III. Materials and Methods

Following [2], our proposed hybrid model consists of two stages. The first stage is clustering using Agglomerative Hierarchical Clustering algorithm. The resulting clusters from this first stage is fed into the second stage, which is classification using Decision Tree J48 Classifier. The model attempts to improve the classification accuracy in predicting patients as diabetic or otherwise based on the patients' medical profiles. Fig. 1 shows the architecture of our proposed hybrid classification model. The following subsections will present in detail on how the model is implemented.

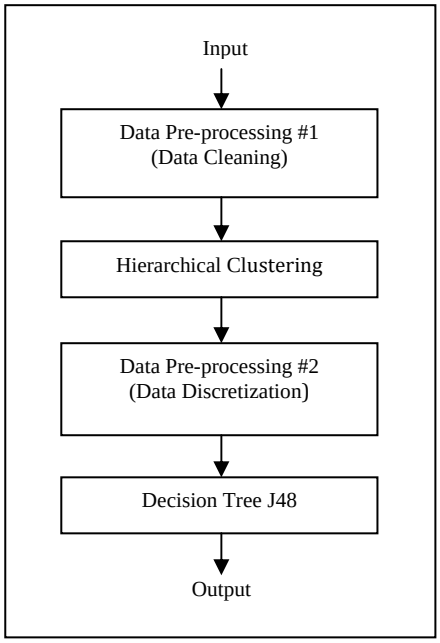

Fig. 1. Architecture of the proposed hybrid model

A. Data Description

The proposed hybrid model will be validated on Pima Indians Diabetes Dataset (PIDD), sourced from UCI Machine Learning Repository [14]. This dataset consists of medical information on 768 female patients of Pima Indians heritage. In particular, the database comprises 8 attributes (all numeric-valued) related to personal and medical features and one class valued 0 (interpreted as "tested negative for diabetes") or 1 (interpreted as "tested positive for diabetes"). Out of 768 instances, 500 patients were tested negative for diabetes and the remaining 268 patients were tested positive for diabetes. Table 2 shows the mentioned 8 attributes including the class for this dataset.

TABLE II
Attributes and Classes for PIDD

Category	Attributes/Class	Abbreviation
Personal	• Number of times pregnant	Pregnant
	• Age (Years)	Age
Medical	• Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Plasma
	• Diastolic blood pressure (mm Hg)	BP
	• Triceps skin fold thickness (mm) • 2-Hour serum insulin (mu U/ml)	Skin Insulin
	• Body mass index (weight in Kg/Height in m)^2	BMI
	• Diabetes pedigree function	Pedigree
	• Class variable (0=Tested Negative or 1=Tested Positive)	Class

B. Data Pre-processing

In reality, data are often incomplete in consequence of deficiency in attribute values, interesting attributes of interest or contain only aggregate data. Data are also susceptible to noise, in particular errors or outliers. In order to have a quality result in data mining, accuracy, completeness, consistency, reliability and accessibility of the data should be considered. The pre-processing activities for this project focus on data cleaning in order to manage missing values and continuous data.

1) Data Cleaning: The aim of data cleaning is to handle the missing values, smoothing noisy data, identifying or removing outliers and accomplish inconsistencies of the data. Furthermore data cleaning might intensify the quality of data to a level suitable for the selected analyses as well. Referring to the Pima Indian Diabetes dataset, there are five attributes with missing values, which are glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, serum insulin, and body mass index. To sort out missing value for those attributes, records that have missing values are deleted. There is one (1) record of outlier being omitted as well.

2) *Data Discretization:* Some data mining algorithms are not capable to handle continuous variables. Meaning that, these types of data should be changed into categorical format in order to proceed with the chosen data mining tasks. This process is called data discretization. In this study, data discretization is performed by dividing the range of continuous attribute into intervals. Each interval labels can then be used to replace actual data values and we binned the data into 3 whereby each containing approximately has a same number of samples. This is essentially equal-depth (frequency) partitioning and data is divided into three samples based on the opinion of medical experts in [3].

C. Hierarchical Clustering

The clustering stage was implemented with agglomerative hierarchical clustering algorithm, which is HierarchicalClusterer in WEKA environment. The process started with all feature points as individual cluster. At each consecutive step, this algorithm merged the closest pair of clusters until there was left only one cluster based on Euclidean distance calculation. This requires defining the notion of cluster closeness.

D. Decision Tree Classification

The classification stage in the proposed hybrid model was implemented using a variant of C4.5 Decision Tree algorithm, which is J48 in WEKA environment. During this stage, the correctly clustered instances from Stage 1 were discretised as served as an input to this classifier. The classification experiment was performed using hold-out validation method with 80% training and 20% testing. The output of this stage is the classification accuracy.

IV. EXPERIMENTAL RESULTS

The performance accuracy for the standard Decision Tree J48 classifier was 76.9%. However, the proposed hybrid model with HierarchicalClusterer and J48 achieved higher accuracy rate of 80.8%. Experimental results showed an improvement in accuracy diabetic dataset using the proposed hybrid model as compared to Decision Tree J48 alone. Table 3 shows the comparison of the results.

TABLE III
Results for the Classification Experiments

Task	**Algorithm**	**Accuracy (%)**
Classification	Decision Tree J48	76.9
Hybrid Model of Clustering and Classification	Agglomerative Hierarchical + Decision Tree J48	80.8

Next, Table 4 shows the confusion matrix where the upper row denotes the result for the negative class and the lower row denotes the result for the positive class.

TABLE IV
Confusion Matrix of the Proposed Hybrid Model

a	***b***	**Classified**
45	7	*a* = Tested Negative
8	18	*b* = Tested Positive

Based on the confusion matrix, a number of other metrics can be derived, in particular the true positive (TP), which represent patients who are correctly predicting as having diabetic and the false positive (FP) which patients who are incorrectly predicting as having diabetic. Using these two values, we can pick up true positive rate (TPR) and false positive rate (FPR). We plotted the result in Receiver Operating Characteristic (ROC) space. ROC is an important method for this study since it identifies the relationship between the FP and TP. As shown in Fig. 2, the plotted result in ROC space shows that the result is plotted at the upper left corner. It indicates that by having hybrid model, the percentage of predicting *n* (negative) and *p* (positive) in having diabetic of the patients is high.

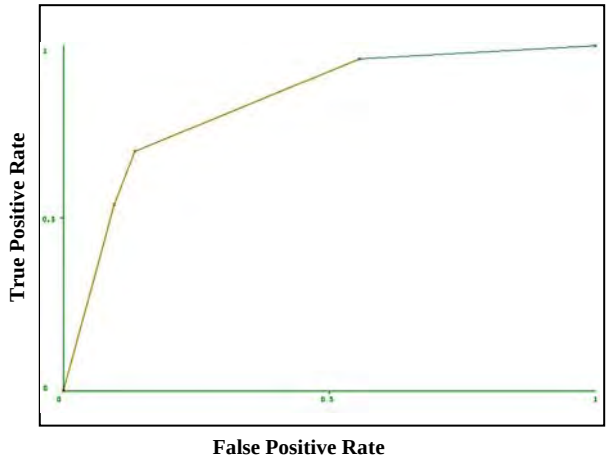

Fig. 2. ROC curve of the proposed hybrid model

In addition for observing further result, the detailed accuracy by class of the proposed hybrid model on FPR, precision and ROC area are shown in Table 5.

TABLE V
Detailed Accuracy of the Class of the Proposed Hybrid Model

FPR	Precision (PPV)	ROC Area	Class
0.31	0.85	0.83	0 = Tested Negative
0.14	0.72	0.83	1 = Tested Positive

Precision is equivalent to positive predictive value (PPV) which signifies the percentage of patients who are truly predicted as *p* or having diabetic (TP/(TP+FP)). Meanwhile, ROC area indicates the area under the ROC curve (AUC) which the value will always be between 0.0 and 1.0. Since the result of ROC area/AUC is 0.8, which is close to 1.0, it indicates high performance accuracy of the proposed hybrid model.

V. Conclusions

Data mining in medicine plays very important role in improving general health and community life at present. The main aim of this paper is yet to improve the performance accuracy of existing hybrid model by incorporating new clustering algorithm into the Decision Tree-based classifier. The result shows that the proposed hybrid model achieved higher accuracy rate with application of Agglomerative Hierarchical Clustering. For future work, we will use the hybrid model to conduct a case study for Malaysian medical profiles.

References

[1] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Rule-based classification for diabetic patients using Cascaded *K*-Means and Decision Tree C4.5," *International Journal of Computer Applications*, 45(12):0975 – 8887, 2012.

[2] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Cascading *K*-means clustering and *K*-Nearest Neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology*, 1(3):2249 – 8958, 2012.

[3] A. Albright, "The public health approach to diabetes," *Am. J. Nurs.*, vol. 107, p.39-42, 2007.

[4] B.S Ilango and N. A. Ramaraj, "A hybrid prediction model with F-score feature selection for type II Diabetes databases," in *Proc. of the 1st Amrita ACM-W Celebration on Women in Computing in India*, 2010.

[5] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufamann Publisher, 2009.

[6] A. Mehmet, I. Cigdem and A. Mutlu, "A hybrid classification method of K Nearest Neighbor, Bayesian Methods and Genetic Algorithm," *Expert Systems with Applications* vol. 37, p. 5061–7, 2010.

[7] I. H. Witten, F. Eibe, and M. A. Hall, *Practical Machine Learning Tools and Techniques*, Burlington: USA, 2011.

[8] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for type-2 diabetic patients," *Expert Systems with Applications* vol. 37, p. 8102–8, 2010.

[9] C. Jin, L. De-lin, and M. Fen-xiang, "An improved ID3 decision tree algorithm," In *Proc. of 4th International Conference Computer Science & Education*, p. 127-30, 2009.

[10] S. Zhang, S. Liu, J. Ou, and G. Wang G, "C4. 5-based classification rules mining of high-rise building SFIO," In *Proc. of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 4, p. 467-72., 2008.

[11] A. Rajput, R. P. Aharwal, M. Dubey, S. P. Saxena, M. Raghuvanshi, "J48 and JRIP rules for e-governance data." *International Journal of Computer Science and Security*, 5(2):201-7, 2011.

[12] K. Humar and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, 3(1):82-9, 2008.

[13] S. Aruna, S. P. Rajagopalan, and L. V. Nandakishore, "An empirical comparison of supervised learning algorithms in disease detection," *International Journal of Information Technology Convergence and Services*, 1(4):81-92, 2011.

[14] A. Frank and A. Asuncion, "UCI machine learning repository," [Online]. Available: http://archive.ics.uci.edu/ml.